他話者의 勵起信號를 이용한 抑揚變換

Intonation Conversion using the Other Speaker's Excitation Signal

이 기 영*, 최 창 석**, 최 갑 석**, 이 현 수**
(Ki Young Lee*, Chang Seok Choi**, Kap Seok Choi**, Hyun Soo Lee**)

요 약

본 논문에서는 원음성을 원하는 억양의 음성으로 변환시켜 주기 위한 기초연구로서 타화자의 여기신호를 이용한 억양변환방법을 제안하였다. 이방법에서는 타화자의 여기신호를 억양정보로 이용하였으며, 타화자의 성도스펙트럼과 DTW에 의해 정합되는 원신호의 성도스펙트럼를 추출하여 여기신호의 스펙트럼과 곱한 후 단시간푸리에 역변환해 줌으로써 억양변환된 음성을 합성하였다. 본 방법에 의해 억양변환된 합성음성을 평가하기 위하여 30명의 남성화자가 발성한 한국어 단모음과 문장음성을 대상으로 억양변환실험을 수행한 후 기본주파수의 궤적과 스펙트로그램 및 왜곡측정을 비교하고 MOS 테스트를 실시한 결과 제안된 방법에 의해 임의의 음성을 타화자음성의 억양으로 변환시킬 수 있음을 확인하였다.

ABSTRACT

In this paper an intonation conversion method is presented which provides the basic study on converting the original speech into the artificially intoned one. This method employs the other speaker's excitation signals as intonation information and the original vocal tract spectra, which are warped with the other speaker's ones by using DTW, as vocal features, and intonation converted speech signals are synthesized through short-time inverse Fourier transform(STIFT) of their product. To evaluate the intonation coverted speech by this method, we collect Korean single vowels and sentences spoken by 30 males and compare fundamental frequency contours, spectrograms, distortion measures and MOS test between the orginal speech and the converted one. The result shows that this method can convert and speech into the intoned one of the other speaker's.

I. 序 論

음성에 포함된 정보를 크게 두가지로 나누면 언어정보와 비언어정보로 나눌 수 있다. 비언어적 정보는 음성의 의미를 전달하는 언어적 정보 이외의 개인성, 정서성 및 심리적인상 등의 여러가지 정보를 의미하며 음성의 비언어적 정보를 변경시켜 주면 개인성에서부터 심리적인상에 이르기까지 변환시켜주는 것이 가능하다[1, 2].

음성의 비언어적 정보를 변환하기 위해서는 주로

*관동대학교 전자통신공학과
Dept. Electr. Comm. Eng. in Kwan Dong Univ.
**명지대학교 정보통신공학과
Dept. Inform. Comm. Eng. in Myong Ji Univ.
접수일자 : 1995년 2월 20일

음성의 억양으로 나타나는 운율정보와 성도스펙트럼으로 나타나는 음향정보를 모두 또는 하나를 선택하여 변경함으로써 가능하다. 여기서, 운율정보로 나타나는 억양을 변경해 주면 동일한 서술어로 구성된 문장음성의 구문형태나 화자의 의도를 변화시켜 줄 수 있으며[3], 특히, 전화안내서비스의 합성음성이나 자동번역기의 기계음성에 개인성과 정서성 및 심리적인 상을 부여한다든가 방송국 및 공상과학영화의 특수음성합성을 위한 기술에 기여할 것이다[4-7]. 이러한 운율정보로 나타나는 억양을 변경하기 위하여 피치주기를 변화시키려는 연구가 진행되어 왔으나 인위적인 변경에 의해 재합성된 음성에 왜곡이 생기고 음질이 떨어지는 문제가 발생하여 단일펄스음원을 이용하는 연구[8]와 영추가/제거[9]에서부터 최근에는 PSOLA[10]와 피치반분법[11] 및 Fant의 모델[12]을 개선하는 등에 의해 재합성된 음성의 음질을 높이려는 연구가 관심을 끌고 있다. 그러나 피치의 변경과정에서 인위적인 피치주기의 변경은 재합성된 음성의 음질을 떨어뜨리는데 결정적인 역할을 할 수 있다.

본 연구에서는 주로 문장음성에 운율정보로 나타나는 억양을 변환시키기 위한 연구로서 타화자의 여기신호를 이용한 억양변환방법을 제안하였다. 이 방법에서는 타화자의 여기신호를 억양성분으로 이용하기 때문에 인위적으로 피치주기를 변경하지 않는 이점이 있으며, DTW에 의해 정합된 원신호의 성도스펙트럼을 추출하여 타화자의 여기스펙트럼과 곱한 후 단시간푸리에 역변환하여 음성을 합성하므로써 원음성의 성도특성을 살리면서 억양을 타화자의 것으로 변환하였다. 본방법에 의해 억양변환된 합성음성을 평가하기 위하여 30명의 남성화자가 발성한 한국어 단모음과 문장음성을 대상으로 억양변환실험을 수행한 후 기본주파수의 궤적과 스펙트로그램을 비교하고 왜곡추정 및 MOS 테스트를 실시하여 억양변환의 정도 및 음질을 평가하였다.

Ⅱ. 抑揚變換의 原理

전형적인 성질변환방법으로 선형예측모델을 이용한 분석/합성기법이 있다. 이 기법은 자연음성을 분석하여 성대의 진동이나 성도에서의 공진등, 음성의 성질을 결정하는 특징파라메타를 각각 독립적인 시계열형태로 추출하여 그 특징파라메타을 변경하여 그에 따른 성질변환된 음성을 합성한다. 그림 1은 본 연구에서 제안하는 억양변환의 원리도를 나타내었다.

즉, 음성의 억양을 변화시키기 위하여 입력음성의 여기신호 e(n)의 피치주기를 변경하고 또 성도의 특징을 변화시키기 위해서는 성도스펙트럼 H(w)의 모양을 변형하여 새로운 성도특성을 구성하며 최종적으로 변경된 여기신호 e′(n)를 변형된 성도스펙트럼 H′(w)와 결합하면 억양변환된 음성파형을 얻을 수 있다.

Ⅲ. 제안된 抑揚變換

임의의 화자가 발성한 원음성의 억양을 원하는 타화자음성의 억양으로 변환시키기 위하여 본 연구에서 제안하는 음성의 억양변환은 먼저 타화자가 발성한 음성의 억양성분으로 여기신호의 스펙트럼을 추출하고, 각 억양성분에 정합되는 원음성의 성도스펙트럼을 생성하며, 단시간 푸리에 역변환에 의해 억양이 변환된 음성을 합성한다. 여기서 여기신호의 스펙트럼 E(w)은 단시간-푸리에변환(STFT)된 스펙트럼 X(w)을 선형예측분석에 의한 성도스펙트럼 H(w)으로 나누어 추출하였으며 이를 식으로 나타내면 다음과 같다.

$$E(w) = X(w)/H(w) \tag{1}$$

여기서, p차 선형예측계수 a_k에 의한 성도스펙트럼은 다음과 같다.

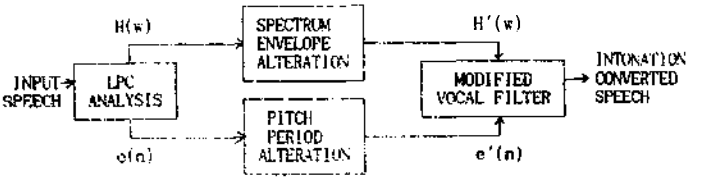


그림. 1. 억양변환의 개념도

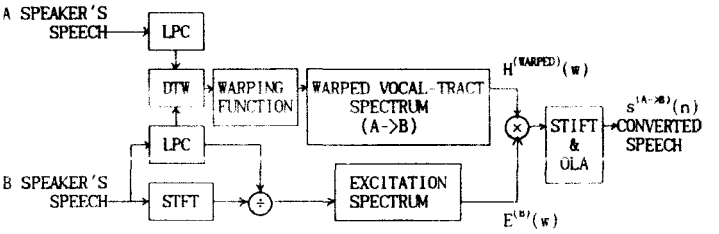


그림 2. 억양변환시스템

$$H(w)=\frac{1}{1+\sum_{k=1}^{p} a_k e^{-jwk}}$$

또한, 정합된 성도스펙트럼은 화자 A와 타화자 B가 발성한 두 음성 사이의 DTW에 의한 정합함수에 의해 결정된다. 그림 2는 A화자의 음성을 B화자의 음성의 억양으로 변환시키기 위한 분석 및 합성시스템을 나타낸 것으로 분석과정은 A화자의 음성에서 B화자의 음성으로 정합되는 A화자의 성도스펙트럼 $H^{(WARPED)}(w)$의 생성과 B화자 음성의 여기스펙트럼 $E^{(B)}(w)$의 추출에 의해 구성되며, 합성은 분석과정에서 얻은 $H^{(WARPED)}(w)$와 $E^{(B)}(w)$을 곱하여 단시간-푸리에역변환(STIFT)함으로써 억양변환된 음성을 합성한다. 이 합성과정을 식으로 나타내면 다음과 같다.

$$s^{(A\rightarrow B)}(n)=F^{-1}[H^{(WARPED)}(w)E^{(B)}(w)] \qquad (2)$$

그림 2의 억양변환을 위한 합성과정에서 분석된 단시간 프레임들의 연속성을 위하여 현재 프레임의 전반부와 이전 프레임의 후반부를 중복가산(OLA)[13]하였다.

Ⅳ. 抑揚變換實驗 및 評價

1. 음성데이타베이스

본연구에서 억양변환에 사용할 음성데이타는 10대 미만의 남성화자 3명과 30대 남성화자 3명이 두번씩 발성한 한국어 기본모음 7개 { "아", "오", "우", "으", "이", "에"}와 6세부터 85세까지 무작위로 분포된 남성화자 30명이 두번씩 발성한 한국어 문장 "안녕하십니까?"로 하였으며, 실험조건은 표 1과 같다. 여기에서 창함수는 단시간푸우리에변환(STFT) 및 선형예측분석과정에서는 rectangular window를 이용하였으며, 중복가산(OLA)과정에서는 Hanning window를 이용하였다. 또한 왜곡척도로 사용한 log likelihood ratio[14]의 식은 다음과 같다.

표 1. 실험조건

A/D data	10kHz sampling, 16-bit
window function	•rectangular : analysis •Hanning : OLA
window(frame) length	256 samples
window shift	128 samples
LPC analysis order	15
distortion measure	log likelihood ratio

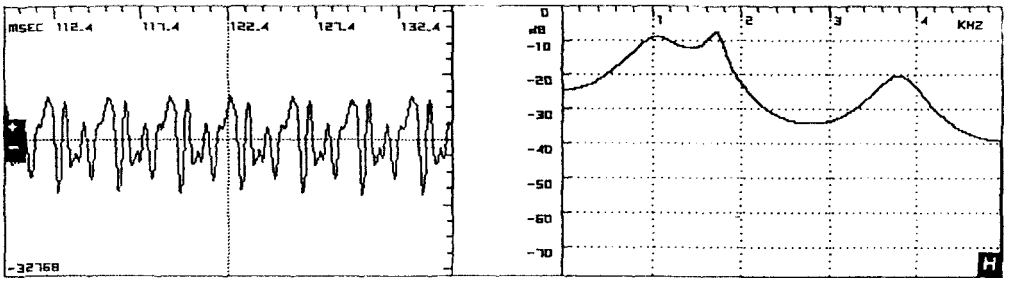
$$d_{ij}=\log\frac{a_j^T R_i a_j}{a_i^T R_i a_i} \qquad (3)$$

여기서 d_{ij}는 i번째 프레임과 j번째 프레임 사이의 왜곡이며 a_i는 i번째 프레임의 선형예측계수의 벡터이고 R_i는 i번째 프레임의 음성신호의 상관행렬이다.

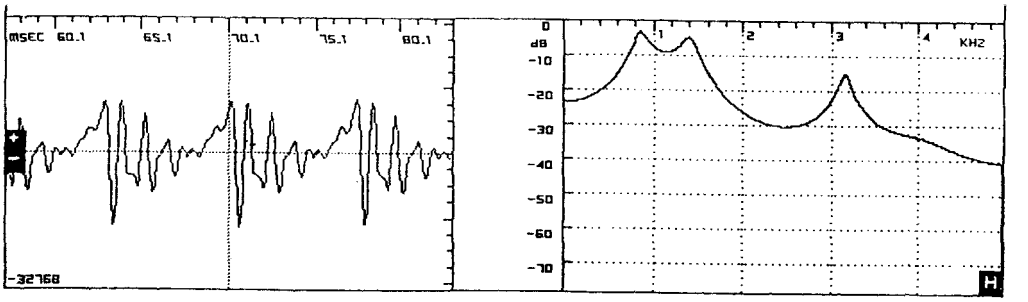
2. 억양변환실험결과 및 고찰

억양은 음성마다 개인차를 뚜렷이 나타내고 있으므로 본실험에서는 단모음을 발성한 화자사이의 억양변환실험과 문장음성을 대상으로 화자사이의 억양변환실험을 수행하였다.

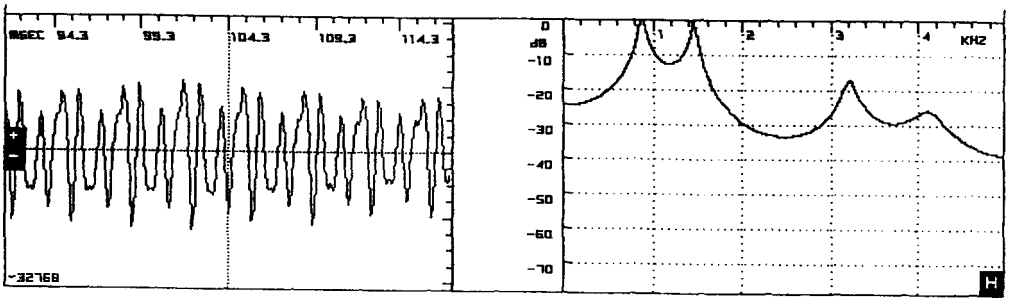
단모음을 대상으로 억양을 변환시키는 경우에는 단모음 일지라도 기본주파수의 변화가 뚜렷하게 나타나는 10대 미만의 남성화자가 발성한 단모음과 비교적 일정한 기본주파수의 30대의 남성화자가 발성한 단모음을 대상으로 실험하였다. 그림3은 10대 미만의 남성화자가 발성한 단모음 "아"와 30대의 남성화자가 발성한 "아" 사이에서 서로의 기본주파수를



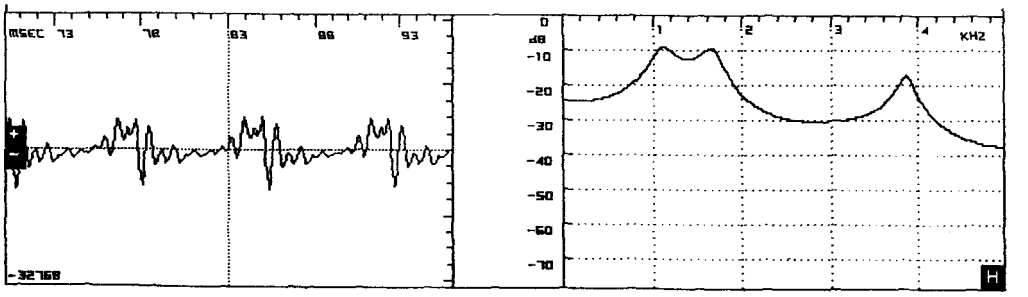
(a) 8세 남성화자의 단모음 "아"



(b) 30세 남성화자의 단모음 "아"



(c) (a)의 기본주파수로 변환된 30세 남성화자의 "아"



(d) (b)의 기본주파수로 변환된 8세 남성화자의 "아"

그림 3. 단모음의 단구간 원음성과 기본주파수를 변환한 합성음성의 비교

변환하였을때 단구간(256ms)의 원음성파형과 변환 합성한 음성파형을 그들의 스펙트럼과 함께 보이고 있다.

그림 3의 (a)원음성의 기본주파수는 284Hz이며 (b)원음성의 기본주파수는 137Hz이다. 또한, 그림 3의 (c)는 식(2)에 의해 (a)음성의 여기스펙트럼과 (b)음성의 정합된 성도스펙트럼을 이용하여 억양을 변환한 음성파형과 그로부터 재추출한 성도스펙트럼이며, 기본주파수는 (a)원음성과 거의 근사한 288Hz로 변환되었으며, 성도스펙트럼도 (b)원음성 성도스펙트럼과 제1, 제2 및 제3포만트주파수까지 동일하다. 그림 3의 (d)는 (c)와 같은 방법으로 (b)원음성의 여기스펙트럼과 (a)원음성의 정합된 성도스펙트럼에 의해 변환한 음성파형과 그로부터 재추출한 성도스펙트럼이다. 여기서, 기본주파수는 (b)원음성과 거의 근사한 136Hz로 변환되었으며, 성도스펙트럼은 (a)원음성의 성도스펙트럼과 제1, 제2 및 제3포만트주파수까지 동일하다. 그러나, 그림 3의 (c)에서 제4포만트주파수가 강조된다거나 (d)에서와 같이 성도스펙트럼의 포만트주파수에 따른 에너지의 크기는 완전히 일치하지 못하고 있다. 그이유는 원음성의 성도스펙트럼과 원음성의 여기신호를 이용하여 재합성하지 않고 타화자의 음성으로부터 여기신호를 추출하여 이용하였기 때문인 것으로 사료된다.

그림 4는 그림 3과 같은 순서로 전구간의 원음성과 변환합성된 음성파형, 스펙트로그램 및 피치궤적을 비교하고 있다. 여기서, 스펙트로그램상의 굵은 곡선은 기본주파수의 궤적이며, (a)와 (c), (b)와 (d)의 각 궤적들이 거의 동일함을 볼 수 있다. 또한, 그림 4의 (a)와 (d), (b)와 (c)의 각 스펙트로그램들이 거의 동일한 모양임을 볼 수 있다. 따라서 이파형으로부터 본연구의 억양변환방법에 의해 원음성이 타음성의 억양으로 변환되었음을 확인하였으며 8세의 남성화자의 기본주파수와 성도특성이 여성화자의 기본주파수와 성도특성과 거의 동일하기 때문에 본연구의 억양변환방법은 성별의 구별없이 적용이 가능할 것으로 사료된다.

문장음성을 대상으로 하는 경우에는 30명의 전화자를 대상으로 억양변환실험을 하였으며 문장음성 사이의 억양변환을 위하여 단모음과 마찬가지로 식(2)를 이용하였다. 그림 5는 "안녕하십니까?"라는 문장음성을 대상으로 억양변환한 합성음성의 일예로

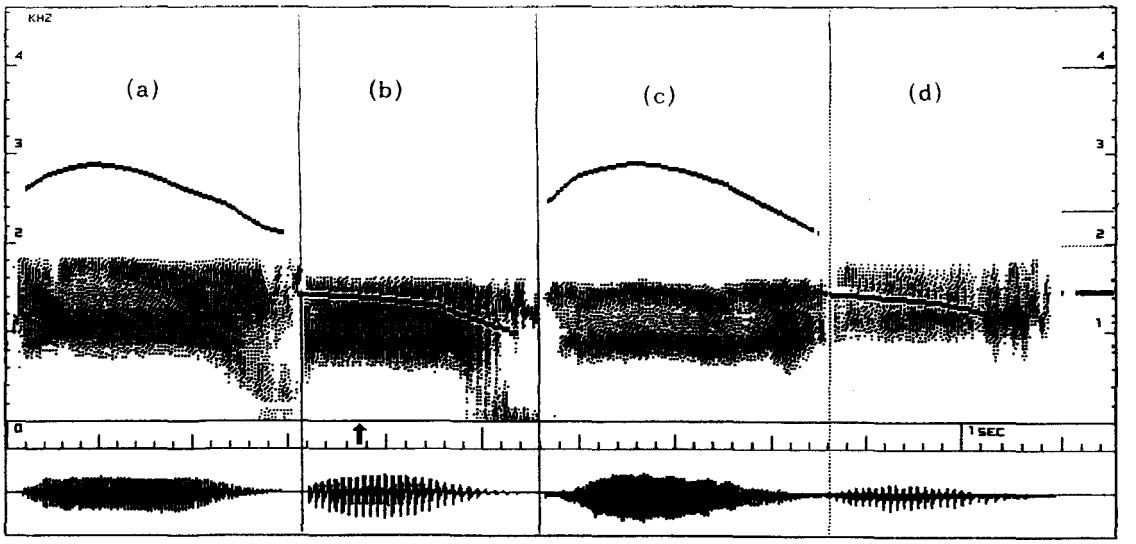


그림 4. 억양변환된 합성음성의 파형과 스펙트로그램 및 기본주파수궤적의 비교
(a) 8세 남성화자의 단모음 "아"
(b) 30세 남성화자의 단모음 "아"
(c) (a)의 억양으로 변환된 30세 남성화자의 "아"
(d) (b)의 억양으로 변환된 8세 남성화자의 "아"

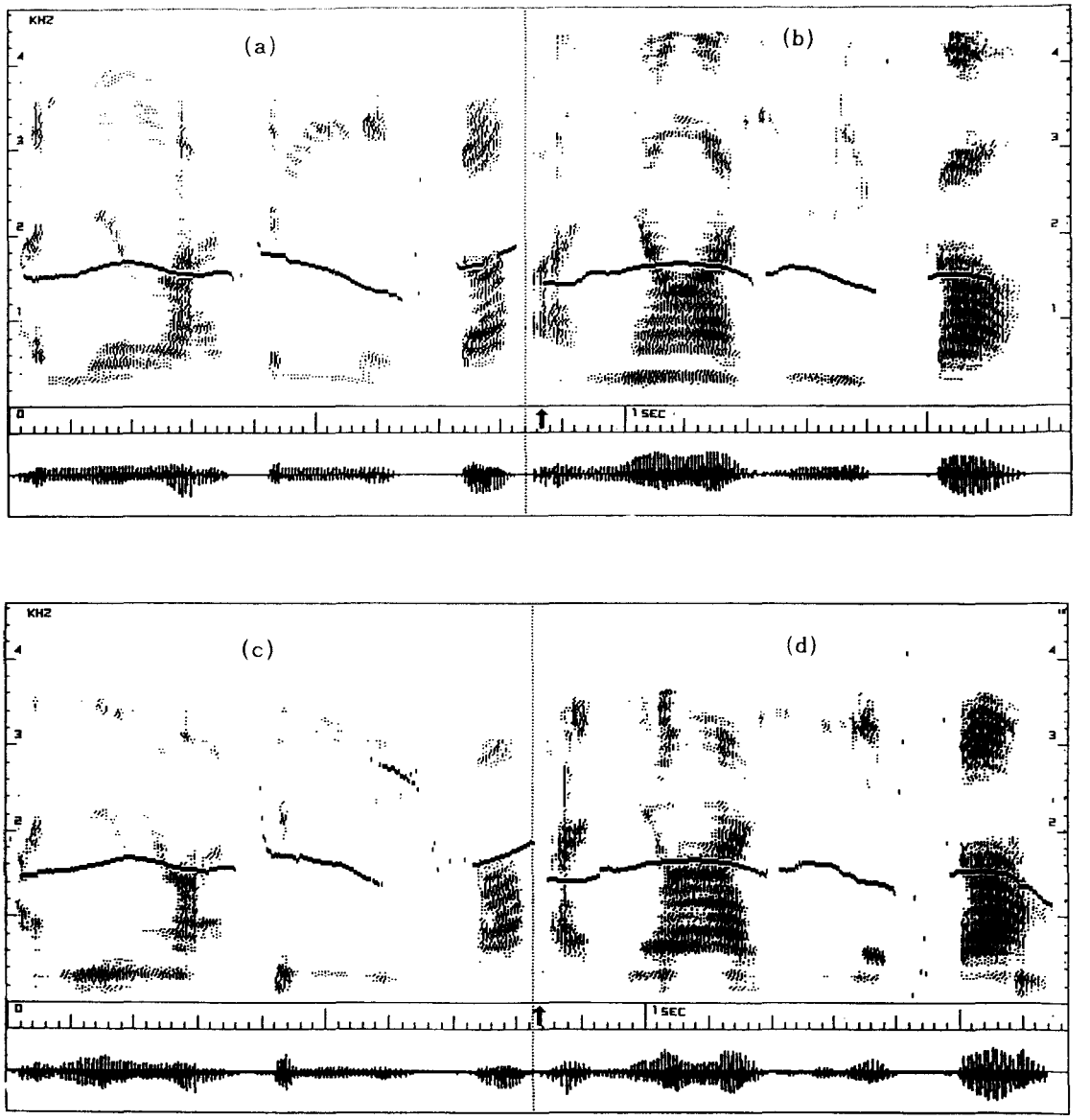


그림 5. A화자와 B화자 사이의 억양변환된 문장음성 "안녕하십니까?"의 파형, 스펙트로그램 및 기본주파수궤적의 비교
(a) A화자의 원음성 (b) B화자의 원음성
(c) A화자의 억양으로 변환된 B화자의 합성음성
(d) B화자의 억양으로 변환된 A화자의 합성음성

문장의 끝을 올리는 화자와 반대로 문장의 끝을 내리는 서로 다른 두 화자사이에 서로 억양변환하여 합성한 실험결과를 보이고 있다.

그림 5에서 (a)와 (c)및 (b)와 (d)의 각 기본주파수궤적들이 거의 동일하므로 문장음성 사이에서도 타음성의 억양을 이용한 원음성의 억양변환이 가능함을 알 수 있었다.

표 2는 억양변환시 발생하는 왜곡을 검토하기 위해

표 2. 왜곡의 비교

비 교 항 목	단모음	문장음성
원신호와 타음성 사이의 왜곡	0.409510	0.777906
원음성과 억양변환된 음성 사이의 왜곡	0.198251	0.286178
타음성과 억양변환된 음성 사이의 왜곡	0.482771	0.476785

억양변환전의 원음성과 타음성 사이의 왜곡, 원음성과 억양변환된 합성음성 사이의 왜곡 및 여기신호를 추출하는 타음성과 억양변환된 합성음성 사이의 왜곡을 측정하여 비교하였다. 여기서 왜곡측정방법은 원음성과의 성도특성을 비교하기 위하여 log likelihood ration로 측정하였다.

표 2에서 원음성과 타음성 사이의 왜곡과 원음성과 억양변환된 합성음성 사이의 왜곡을 비교하면 단모음에서는 약 48%, 문장음성에서도 약 37% 정도 왜곡이 감소됨을 알 수 있으며, 타음성과 억양변환된 합성음성 사이의 왜곡을 비교하면 단모음에서는 오히려 증가된 왜곡이 측정되고 문장음성에서도 원음성과 억양변환된 합성음성 사이의 왜곡보다 큰 왜곡이 측정되었다. 이상과 같이 왜곡을 비교한 결과 억양변환된 음성의 성도특성이 원음성과 가깝도록 합성되었음을 알 수 있었다.

3. 억양변환 합성음성의 평가

실험결과로부터 본연구에서 제안한 억양변환에 의해 문장음성을 합성하였을때 억양변환의 정도와 음질을 평가하기 위하여 MOS 테스트로 비교하는 실험을 하였다. MOS 테스트의 실험대상은 대학교의 연구자 15명으로 하였으며, 원음성과 억양변환된 합성음성을 각각 청취시킨 후, 억양변환이 되었으면 1 아니면 0으로, 자연스럽게 들리며 1 아니면 0으로 또한 명료하게 들리면 1 아니면 0으로 하여 누적하였고, 이누적결과들을 배분율로 표현하여 표 3에 나타내었다.

표 3에서 본논문의 억양변환방법에 의해 원음성의 억양변화는 그림 4와 5에서 추출된 기본주파수의 궤적에서도 확인한 바와 같이 원하는 화자의 억양으로 100% 수행되어 짐을 알 수 있었다. 음질을 평가하기 위해 원음성과 억양변환된 합성음성을 청취시켰을 때 자연성은 93%인 반면, 명료성은 53%인 결과를 보이고 있다. 이것은 그림 3의 (c)와 (d)에서와 같이 합성음성에서 재추출한 성도스펙트럼에서 각 포만트 성분이 분명히 나타나며 억양은 타화자의 것을 그대로 이용했기 때문에 100% 억양변환된 합성음성의 자연성이 비교적 높다고 사료되는 반면, 명료성이 자연성보다 비교적 낮은 것은 "안녕하십니까?"의 억양변환된 합성음성을 청취할때 다른 부분의 음성은 명료한데 비하여 "안녕"의 "초성ㄴ"에 해당하는 부분이 원음성에 비해 약간 불분명하다는 의견이 15명중 7명의 MOS 테스트의 청취자들에게서 있었기 때문이다. 이것은 타화자의 음성에 대한 원음성의 정합된 성도스펙트럼을 DTW의 정합함수에 의해 결정할 때 동일화자 사이에서는 음성의 발성길이가 서로 달라도 동일음소끼리 정합시킬 수 있지만 정합시킬 두 음성의 화자가 다르면 각화자마다 그성도특성이 다르므로 두 화자가 각각 동일한 음소를 발성한다고 할찌라도 DTW의 왜곡측정에 의해 타화자의 억양으로 정합될 성도스펙트럼을 일치하는 음소의 위치에 정확히 구해내지 못했기 때문인 것으로 판단된다.

표 3. MOS 테스트의 결과

기 준	억 양	자연성	명료성
변분율	100%	93%	53%

V. 結 論

본 연구에서는 임의 화자가 발성한 원음성의 억양을 타화자의 것으로 변환시키는 억양변환방법을 제안하였으며, 본방법에 의해 억양변환된 합성음성을 평가하기 위하여 단모음과 문장음성을 대상으로 억양변환을 위한 분석 및 합성실험을 수행한 결과 다음과 같은 결론을 얻을 수 있었다.

1. 타화자의 여기신호를 이용하여 억양변환이 가능함을 확인하였다.
2. 30세의 남성화자의 음성을 기본주파수가 여성화자와 거의 같은 8세의 남성화자의 억양으로 변환하는 합성이 가능하므로 본연구의 억양변환방법이 성구별없이 적용이 가능함을 알 수 있었다.
3. 억양변환된 합성음성의 음질을 평가하기 위하여 MOS 테스트를 실시한 결과 일부를 제외하고 전체적으로 문장음성의 자연성과 명료성이 높은 것으로 확인되었다.

억양변환된 합성음성이 보다 높은 음질을 유지하기 위해서는 정합된 성도스펙트럼을 구하기 위한 DTW 과정에서 화자가 다르더라도 동일한 음소끼리 정합

될 수 있도록 TL 과정을 개선한다면 가능할 것으로 사료된다.

〈본 논문은 한국과학재단 핵심전문연구(과제번호 931-0900-039-2)의 지원에 의하여 이루어졌음.〉

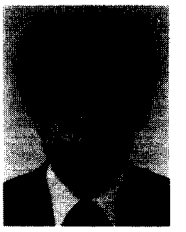
References

1. T.Takagi, T. Umeda, "Voice quality conversion with correction of spectral distortion by pitch manipulation, and its subjective evaluation," the Transactions of the Institute of Electronics, Information and Communication Engineers A Vol. J73-A, No. 3, pp. 387-396, Mar. 1990
2. T.Takagi, "Voice quality conversion," the Trans. Televison, Vol. 47, No. 12, pp. 28-32, 1993
3. 이기문, 외 2인, 국어음운론, 학연사
4. M.A.Jasiuk, V.Geocharoff, J.N.Damoulakis, "Improved speech modification method," ICASSP 87, pp. 1465-1468, 1987
5. D.W.Griffin, J.S.Lim, "Signal estimation from modified short-time fourier transform," IEEE Trans. Acoust., Speech Signal Processing, Vol. ASSP-32, No. 2, pp. 236-243, Apr. 1984
6. T.F.Quatieri, R.J.McAulay, "Speech tranformations based on a sinusoidal representation," IEEE Trans. Acoust., Speech Signal Processing, Vol. ASSP-34, No. 6, pp. 1449-1464, Dec. 1986
7. M.Abe, S.Nakamura, K.Shikano, H.Kuwabara, "Voice conversion through vector quantization," ICASSP 88, pp. 655-658, 1988
8. B.S.Atal, S.L.Hanuauer, "Low-bit-rate speech transmission by linear prediction of speech signals," J.Acoust.Soc.Am. 49, 133(A), 1971
9. B.E.Caspers, B.S.Atal, "Changing pitch and duration in LPC synthesized speech using multipulse excitation," J.Acoust. Soc. Amer., suppl., Vol.73, No.1, pp.S5, Spring, 1983
10. H.Valbret, E.Moulines, J.P.Tubach, "Voice transformation using PSOLA Technique," EUROSPEECH 91, pp. 345-348, 1991
11. 배명진, "고음질합성을 위한 피치변경법," 한국음향학회지 12권2호, 1993
12. G.Fant, J.Lilijencrants, Q.C.Lin, "A four-parameter model of glottal flow," STL-QPSR 4/1985, pp. 1-13, 1985
13. F.J.Charpentier, M.G.Stella, "Diphone synthesis using overlap-add technique for speech waveforms concatenation," ICASSP 86, pp. 2015-2018, 1986
14. A.H.Gray, J.D.Markel, "Distance Measures for Speech Processing," IEEE Trans. Acoust., Speech Signal Processing, Vol. ASSP-24, No. 5, Oct. 1976

▲이 기 영 1961년 5월 7일생

1984년 2월 : 명지대학교 전자공학과 졸업
1986년 2월 : 명지대학교 전자공학과 석사과정 졸업 (공학석사)
1992년 2월 : 명지대학교 전자공학과 박사과정 졸업 (공학박사)
1993년 3월~현재 : 관동대학교 전자통신공학과 조교수

▲최 창 석 1954년 7월 15일생



1978년 2월 : 홍익대학교 전자공학과 졸업
1988년 2월 : 일본 金澤(가나자와) 대학원 전기정보공학과 석사과정 졸업 (공학석사)
1991년 2월 : 일본 金澤(가나자와) 대학원 전기정보공학과 박사과정 졸업 (공학박사)
1984년 1월~1992년 2월 : 산업기술정보원 책임연구원
1993년 3월~현재 : 명지대학교 정보통신공학과 조교수

▲최 갑 석 : 12권5호 참조

▲이 현 수(Hyun Soo Lee) 1951년생

1974년 : 서울대학교 전자공학과 졸업(공학사)
1983년 : 프랑스 대장성대학교 석사학위 취득
1992년 : 프랑스 르아브르대학 박사학위 취득
1992년~현재 : 명지대학교 조교수
※관심분야 : 영상처리, 영상인식